Money, lies, and replicability: On the need for empirically grounded experimental practices and interdisciplinary discourse

Ralph Hertwig & Andreas Ortmann*

Authors' response
(to the commentaries on "Experimental practices in economics: A methodological challenge for psychologists?")

Abstract

This response reinforces the major themes of our target article. The impact of key methodological variables should not be taken for granted. Rather, experimental practices should be grounded in the empirical evidence. If no evidence is available, decisions on design and implementation should be subjected to systematic experimentation. In other words, we argue against empirically blind conventions and against methodological choices based on unreflected beliefs, habits or rituals. Our approach will neither inhibit methodological diversity nor constrain experimental creativity. More likely, it will promote both goals.

Corresponding authors:

Ralph Hertwig
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany
E-mail: hertwig@mpib-berlin.mpg.de
rhertwig@paradox.psych.columbia.edu

Andreas Ortmann
Center for Economic Research and Graduate Education
Charles University
Politickych veznu 7, 111 21 Prague 1, Czech Republic
E-mail: andreas.ortmann@cerge.cuni.cz
aortmann@yahoo.com

Table of Contents

1. Introduction

We concluded the target paper in the hope that it would "spur psychologists and economists to join in a spirited discussion of the benefits and costs of current experimental practices." We are delighted to see that our hope has become reality. Psychologists and economists, together with researchers from other disciplines, responded to our "gentle aggression" **(Harrison & Rutström)** by contributing to an overdue conversation on the nature, causes, and consequences of the diverging experimental practices in psychology and economics. We would be pleased if this discourse indeed moved us "closer to a common language for scientific discovery" **(Harrison & Rutström)**.

Our reply encompasses two major sections. In the first section, we address commentators' responses to our discussion of four key variables of experimental design. The second section is organized around a number of additional issues - among them are conjectures on the causes of the methodological differences and the affinity between experimental economics and behaviorism. We conclude with an outline of some aspects of experimentation in psychology from which we believe economists could learn. We begin with a discussion of what appears to be the most serious and common concern with regard to our analysis.

2. Do our policy recommendations jeopardize experimental diversity and creativity?

A number of commentators (e.g., and prominently, **Gil-White, Guala, Hilton, Huettel & Lockhead**, and **Kurzban**) argue that our recommendations - to ground design decisions in empirical evidence, to systematically manipulate key variables of experimental design (as expressed by our do-it-both-ways rule[1]), and to use deception as practice of truly last resort - would stifle methodological diversity and constrain experimental creativity. Both are important goods - endangering them would compromise our policy recommendations. Like **Maratsos** and **Harrison & Rutström**, however, we believe that our recommendations will not have this effect. Rather, we anticipate our recommendations would promote both experimental diversity and creativity. To explain why, we first summarize our argument and - drawing on commentators' objections - refine it.

We documented in the target article that many experimenters, in the field of behavioral decision making and related areas in social and cognitive psychology (i.e., areas of interest to both psychologists and economists; see Rabin 1998), tend to realize key variables of experimental design in a fast (e.g., using snapshot studies, and brief scripts), inexpensive (e.g., offering no financial incentives), and convenient (e.g., using deception) way. The drift toward these and other seemingly cost-effective experimental methods such as large classroom studies and take-home questionnaires **(Gigerenzer)** occurred, so we argue, due to a lack of strong conventions and a theoretical framework which suggest how to implement experimental tests. While it is rational for experimenters as individuals to select methods and evolve conventions that minimize the costs (in time and money) of producing publishable data, we documented that this preference has a price tag too often overlooked: a greater likelihood of systematic data variability and error variance than alternative (and more expensive) methods. Ultimately, the predominance of fast,

inexpensive, and convenient methods of data collection is likely to contribute to a lack of replicability of experimental results. We identify the fast, inexpensive, and convenient route to data collection as the first source of data variability. Experimental practices that contribute to this first source of data variability undermine control - "the hallmark of good experimental practice, whether it be undertaken by economists or psychologists" **(Harrison & Rutström).**

There is a second source of data variability in psychological experimentation, namely, variability due to methodological diversity. Methodological diversity is high in the research areas we focused on because in these areas some researchers choose to implement more "expensive" realizations of the key variables, employing repeated trials and financial incentives and never using deception. The fact that many researchers use seemingly cost-effective methods, whereas others do not, is likely to be a second source of systematic data variability. The variability in empirical findings which we asserted and documented thus draws on two sources: the variability in results due to fast, inexpensive, and convenient experimental methods (what **Davis & Durham** call lack of "reliability" and what **Guala** calls lack of "clear-cut design") and due to the fact that a small but significant number of experimenters actually use other methods (what **Guala** calls "varied designs"). In the target article, we did not distinguish as clearly between these two sources of variability as with thanks to our commentators' insights and our own hindsight we now realize we should have. This unfortunate fact seems to have been the reason for some commentators' concern that we are out to stifle experimental diversity and creativity.

The do-it-both-ways rule (which accords key variables of experimental design the status of independent variables) does not pose a risk to methodological diversity and experimental creativity for three reasons. First, the rule is tailored to the four key variables in question, and is not meant to interfere with other aspects of experimentation (i.e., our discussion has no bearing on the "participant-observer methodology or single case studies in clinical psychology" as **Davis & Durham** suggest). Second, in contrast to **Davis & Durham's**, **Gil-White's,** and **Maratsos's** explicit reading and other commentators' (e.g., **Betsch & Haberstroh, Guala, Suleiman**) implicit suggestion, we do not endorse empirically blind rules such as economists' strict convention of always using financial incentives.[2] Rather, design and implementation decisions ought to be informed by the evidence rather than by unreflected beliefs, habits or rituals. Third, the do-it-both-ways rule - applicable when evidence is unavailable or mixed - is a systematic reminder to implement more than one realization of a key design variable. It acknowledges that methodological variables imply auxiliary hypotheses (**Gillies & Rigdon, Smith**) and makes them an explicit part of theory testing. The do-it-both-ways rule broadens our experimental inquiry as it adds to researchers' methodological repertoire of fast, inexpensive, and convenient methods, alternative realizations of key variables (a consequence that **Maratsos** also foresees). Ultimately, the do-it-both-ways rule will counteract the de facto hegemony of the seemingly cost-effective methods which presently contribute to what we identified as the first source of data variability.

We admit that our suggestion to eschew deception, whenever possible, imposes constraints. We do not think, however, that such a convention undermines experimental ingenuity. The fact that deception - notwithstanding the APA's admonition to use it only as a last-resort strategy- is still frequently used, indicates that there are no strong incentives to develop, evaluate and employ alternatives. Making deception into a strategy of truly last resort is likely to spur the invention of

new methods (as suggested by Baumrind 1971, and as exemplified by Bardsley 2000). We now turn to commentators' responses to our discussion of four key variables of experimental design

3. Enacting a script versus "ad-libbing"

There were apparently misunderstandings and questions about what we meant by a script. Economist **Roth**, for example, was "not sure that the use of scripts is common in experimental economics". And psychologist **Baron**, for example, said, "I do not understand what is not a script ... ." Again: What is a script, and why does it matter?

We defined a script as clear and comprehensive instructions which detail players (e.g., buyer, seller, market analyst, proposer, responder), their action choices, and the possible consequences of their choices (i.e., the payoffs). In addition, we described the particular kind of role-playing typically employed in economics experiments. Letting participants take on a particular role - having them enact a script - can be used to study not only strategic interactions but judgment, reasoning, and memory performance (e.g., Wason selection task, hindsight bias).

In our opinion, having participants enact explicit and comprehensive scripts has four potential advantages. First, scripts may constrain participants' interpretations of the situation by focusing their attention on those aspects that are intentionally communicated by the experimenter. The hindsight bias studies we described illustrate this point. Davies (1992) told participants only to recreate a previous state of knowledge, thus leaving participants to decide whether they should (1) attempt to retrieve their previous judgment as accurately as possible, (2) look as good (i.e., knowledgeable) as possible, or (3) spare their cognitive effort as their recollections would have no tangible consequences. In contrast, in Camerer's et al. (1989) market experiment, the objective of avoiding the hindsight bias followed from being a successful trader; in other words, the role per se clarified the demand characteristic of the situation. Second, scripts can promote participants' active involvement in the experiment by making their choices have tangible consequences. Third, scripts (especially if they are not abstract), and assignment of perspectives, may cue specific inference mechanisms (e.g., Cosmides & Tooby's, 1992, cheating detection algorithm) that otherwise would not be activated.

Finally, explicit and comprehensive scripts are the basis on which the sometimes subtle influence of instructions can be studied. There is, for instance, intriguing evidence that tiny procedural differences can make a large difference to behavior. Recently, Burnham, McCabe and Smith (2000) and Hoffman, McCabe and Smith (2000) showed, for instance, that changing the word "opponent" to "partner" or prompting players to think strategically before making an offer can have significant effects on how they behave in various contexts (for another striking example, see Harrison, 1999, pp. 26-28). Therefore, explicit and comprehensive instructions enhance procedural regularity and ultimately, we claim, replicability (see also Binmore 1999). Scripts are thus one key to understanding the variability of experimental results - in economics and psychology. For an example, consider the test of the ultimatum game which one of our commentators, **Henrich**, conducted.

Whereas previous tests demonstrated that the "normative" solution of an uneven split was not a good description of empirical results, Henrich (2000) found that the Machiguenga people of the Peruvian Amazon make decisions which are much closer to the game-theoretic prediction. Henrich used instructions ("a set script"), but he had to explain the game at least three times. In addition, "often numerous examples were necessary to make the game fully understood" (p. 975). The experiment itself was introduced to an initial group of participants "under the auspices of 'playing a fun game for money'" (p. 975). Whereas Henrich (2000) suggests that "procedural differences seem unlikely to explain the substantial differences in observed behavior" (p. 975), Burnham's et al. (2000) and Hoffman's et al. (2000) results suggest the opposite. Certainly, the specific aspects of design mentioned above represent significant departures from the standard scripting of ultimatum games. For example, there is no telling what the impact was of the repeated explanations of the set-up or the numerous (possibly unscripted) examples. There is also a good chance that the framing of the experiment as a "fun game" being played for money had an impact on the result. While there are still other possible explanations for the surprising results (e.g., the relative social distance among Machiguenga families), we argue that a clear and comprehensive script, to which the experimenters religiously adhered, would have increased one's confidence in the robustness of the reported results.

In the context of scripts, **Suleiman** points out that the "clarification of the demand characteristics ... is unavoidably entangled with the enhancement of those demand characteristics which coincide with the experimenters' focus." We agree and share his concern - as our comment on the reiterated explanations to participants of the ultimatum game in Henrich (2000) illustrates. We also share Suleiman's assessment of a problematic instruction from a well-known economics experiment published in 1985 - although the really interesting issue related to this two-stage implementation of the ultimatum game is, as one of the authors of the incriminated sentence has pointed out elsewhere, "why did (participants) not do what they were told at the first trial?" (Binmore 1999, F20). That said, we suggest that the benefits from clear and comprehensive instructions typically outweigh the costs of demand effects. Of course, even this is an empirical issue and can be approached as such. Relatedly, **Goodie** makes the excellent point that "one does not know if scripted interactions are representative of non-scripted ones." To which we say, amen, and would add that one also does not know whether non-scripted interactions are representative of that which an experimenter would like them to be representative of. Again, even that is an empirical issue.

None of this implies (as, for example, **Davis & Durham** and **Huettel & Lockhead** intimate) that we try to exorcize incomplete information or uncertainty (including expectations regarding the behavior of other participants) from experimental settings. Economists routinely and systematically try to understand how "a choice by one person is affected by her or his expectations about the choices that might be made by other people"(Huettel & Lockhead). As a matter of fact, the recent literature on various bargaining games (e.g., Rabin 1993; Dufwenberg & Kirchsteiger 2000; Falk & Fischenbacher 1999; Charness & Rabin 2000) is all about this issue. Our point is that scripts can help to reduce unwanted, uncontrolled, and unnecessary uncertainty by channeling participants' interpretations of the experimental situation.

To conclude: Of the four key variables of experimental design, the effects of scripts (and their enactment) are most difficult to analyze due to scripts being rarely treated as independent

variables. Thus, the evidence for our claim that scripts affect results is tenuous. While psychologists and economists are likely to share the view that supplying clear and comprehensive instructions is good experimental practice, there is nevertheless a difference between a script that details a role and scant instructions which make no reference to a role or a perspective (which, to answer **Baron**, counts as no script). Does the difference matter? We delineated four reasons why it might matter. Fortunately, whether it actually does can (and, as we argue, should) be investigated experimentally. As **Harrison & Rutström** correctly state, our main suggestion is indeed to encourage the replication of previous designs and to study the effects of orthogonal variations (whether it is scripts or other key methodological variables).

4. Repeated trials versus snapshot studies

No doubt, the study of both one-shot and repeated game and decision situations is useful (see Barron & Erev 2000). Notwithstanding their interest in equilibrium behavior, economists have recently contributed models of (noisy) introspection for one-shot games (namely, Goeree & Holt 1999, 2000, forthcoming; Olcina & Urbano 1994; Stahl & Wilson 1995). What makes these models intriguing is that they provide one explanation for departures from normative solutions Our point is not that there is no place for experiments only carried out once. Rather, our point is that there has been almost no place for repetition and feedback in psychological research (as our analysis of the Bayesian reasoning literature illustrated). Why do we think that repetition and feedback are important?

4.1. Why repetition? Some commentators misunderstood our argument. We advocate repetition not only because the environment typically forces us into repeated decision and game situations (as suggested by **Betsch & Haberstroh**). Rather, our concern is that participants have a chance to become familiar with what is even under the best of circumstances an unusual situation. We gave these "practice effects" (**Baron**) as the first and foremost reason (for a similar argument see Binmore 1999). In decision situations (i.e., "games against nature"), the particular kind of repeated trials which we discussed - stationary replication - means repeated decision making or judgments in the same scenario. In game situations, stationary replication takes the specific form of a "turnpike design" where one makes repeated decisions but encounters the same player(s) only once. A second motivation for the use of repeated trials is specific to interactive situations. Repeated trials of this kind afford participants the opportunity to learn how their own choices interact with those of other players in a specific situation. We acknowledged that these two kinds of learning may be difficult to distinguish. Still, in the target article we pointed out "the first kind of learning (adapting to the laboratory environment) relates to a methodological concern that participants may not initially understand the laboratory environment and task, whereas the second kind of learning (understanding how one's own choices interact with those of other participants) relates to the understanding of the possibly strategic aspects of the decision situation".

4.2. Stationary replication and other repeated trials. **Gillies & Rigdon** take us to task for our rationalization of the frequent use of repetition in economics. They argue that we get the history of game theory backwards. However, we did not aspire to tell the history of the eductive and the evolutive approach to equilibrium (selection) – one of us has done this elsewhere (Goodie et al. 1999) and has contributed to an analysis of the comparative advantages of the two approaches

(Van Huyck et al. 1995; see also Blume & Ortmann 2000) - an issue that we consider to be one of the fascinating issues in economic theorizing. Our point was simply that most economists are interested in equilibrium behavior and that experimentalists often justify their focus on play in late rounds in this manner (Camerer 1997). We did not say that we subscribe to such a research agenda.

Gillies & Rigdon also suggest that we do not understand that repeated games may generate additional equilibria. To repeat, we did not focus on what Gillies & Rigdon call "repetition with replacement". Rather, we discussed stationary replication. In this context (i.e. when one employs a turnpike design), their objection (see also **Harrison & Rutström**) that repetition of the trust game was likely to lead participants to "update their estimates of the distribution of player types (trusters or not) in the environment" is well-taken. However, if experimenters are concerned about such an effect (as well they should be) they can always refrain from giving feedback until all rounds are played. It is possible that this would not get rid of the problem completely because there is evidence that simple repetition even without feedback has effects (e.g., Keren & Wagenaar 1987, or more recently Barron & Erev 2000) but we believe that simple repetition without feedback significantly attenuates the problem brought up by **Gillies & Rigdon** and **Harrison & Rutström**.

Henrich argues that another problem with repeated-game experiments is the almost complete emphasis on studying individual learning, as opposed to social learning. This, lest we misunderstand, strikes us as an untenable statement. Experimental economists routinely study the emergence of conventions (e.g., Young 1993, Van Huyck et al al. 1995). In fact, what is now called (evolutionary) game theory (e.g., Weibull 1995; see also Binmore 1994 and earlier Smith 1976 [1759]) is all about social learning. More generally, as **Ross** points out, so is much of game theory due to its interactive nature.

5. Financial incentives versus no incentives

In response to our analysis of financial incentives, commentators focused on three major issues: the conditions under which incentives are (not) suitable; the difference between real and hypothetical payoffs; and the effect of financial incentives on anomalies (i.e., violations of normative standards) in individual decision making. We address each of these issues in more detail.

5.1. When should incentives be used? We suggested two **criteria** for the use of incentives: that research focus on people's maximal performance and that standards of optimal behavior be available. In addition, we proposed a simple decision tree to determine whether or not incentives should in fact be used when both criteria are met. First, is there evidence in past research regarding the effects of incentives? If "yes", does the available evidence indicate that financial (or possibly other) incentives affect behavior? If "no", we suggested applying a simple do-it-both-ways rule, thus according financial incentives the status of an independent variable.

In light of this approach, we argue warnings that relying exclusively on financial incentives would oversee some important phenomena (**Roth**), or that investigating behavior such as child rearing

using financial incentives would be inappropriate (**Davis & Durham**) as orthogonal to the policy we propose. Evidently, our policy does not adopt economists' current practices lock, stock and barrel, nor does it define financial incentives to be the norm in decision experiments (as suggested by **Gil-White**). Moreover, the policy does not deny the exciting possibility that less effortful processes can outperform more effortful ones (**Betsch & Haberstroh;** see also Hertwig & Todd in press) or that decision parameters differ across domains (**Rakow**). Of course, this approach also does not deny that incentives other than money may motivate participants (e.g., credit points; **Goodie**). In this context, it is heartening to see that even economists have started to explore the effects of financial incentives systematically, rather than taking them for granted. Schotter and Merlo (2000), for example, have translated the insight that exploring the strategy space may be (cognitively) expensive to participants in an experimental design which demonstrates that not paying participants while they learn can lead to significant improvements in outcomes (see also Berninghaus & Ehrhart 1998).

Betsch & Haberstroh discuss a set of four 'principles' that must be met for financial incentives to be beneficial in terms of people's performance. The first 'principle', availability of exact performance criteria, reflects common wisdom. Next, as the second and third 'principles', they stress that 'external' (experimenter's) and 'internal' (participant's) criteria, together with participant's and experimenter's representation of the task, should be congruent. We understand that Betsch & Haberstroh do not presume such congruence to be necessary for financial incentives to be employed (as we can never a priori ascertain it and because payoff instructions are one key route to aligning experimenters' and participants' task representations) but rather that discongruence is a good candidate explanation when financial incentives have no effect or impair performance. As a fourth 'principle' for incentives to be beneficial, Betsch & Haberstroh propose that a deliberate strategy must be the most appropriate mean to solve the problem. They claim that people often use other than deliberate processes to their advantage. This is a point well-taken. Previous research (in psychology and even more so in economics) has underestimated the role of simple, non-effortful, possibly automatic processes. That said, aspects of our memory, judgment and decision-making performance are, as we documented, under strategic control, and thus the amount of cognitive effort invested can affect the performance. But even when less effortful processes are the object of study, the use of financial incentives makes the demonstration that those processes can outperform effortful ones even more convincing. Moreover, they help to explain where and how effortful processes can go wrong. Betsch & Haberstroh's conclusion that because financial incentives "might have counterproductive effects", they ought not to be used is a nonsequitur. To rephrase Betsch & Haberstroh's conclusion, not taking the role of methodological key variables for granted but subjecting them to systematic variation paves the way to good experimentation (see **Fantino & Stolarz-Fantino, Guala, Harrison & Rutström**).

5.2. Aren't hypothetical incentives sufficient? **Kühberger** offers an intriguing argument: Since decision making involves anticipation of hypothetical events (e.g., future feelings, states of the world), hypothetical decisions are a valid proxy for people's real decisions. The implication of his argument is that hypothetical payoffs may suffice to study the average response (albeit we note that financial incentive have been documented to reduce variance (e.g., Smith & Walker 1993; Camerer & Hogarth 1999; Rutström 1998). Kühberger, however, qualifies his claim: The decision's importance can turn hypothetical decisions into invalid proxies for real ones. **Holt &**

Laury demonstrate such a case precisely. In a choice between lotteries, they find comparable amounts of risk aversion for hypothetical and low real payoff conditions. A high real payoff condition, however, produced drastically different risk attitudes. Holt and Laury (personal communication) have since strengthened these results by conducting additional sessions with even higher real payoffs. Relatedly, **Fantino & Stolarz-Fantino** suggest that the importance of what is at stake may also explain why experiments with humans obtain less impulsive behavior than experiments with pigeons. The incentives experimenters offer to pigeons dwarf those offered to humans. (Of course, even here Harrison's, 1994, dominance critique of payoffs might apply.) The domain of high-stake decisions is not the only one where real and hypothetical incentives can yield divergent results. The results reported in Table 2 (of the target article) and in Camerer and Hogarth (1999) demonstrate that payoffs need not be high stake incentives to affect people's judgement and decision making. .

Although **Kühberger** acknowledges that what is at stake matters and **Holt & Laury** demonstrate that high payoffs can cause dramatic differences, they arrive at opposite policy recommendations. The former stresses the need for a theory of when and why real and hypothetical decisions coincide. In the absence of such a theory, he considers the do-it-both-ways rule a waste of money. In contrast, Holt & Laury argue that even while the discipline lacks an accepted theory of when financial incentives matter, they should nonetheless be used in economics experiments. If they mean there should be a religiously enforced convention to always use financial incentives to manipulate this variable orthogonally, we disagree. If, however, their argument is intended as a call to manipulate the provision of incentives systematically (as they in fact did), we agree with them. In contrast to Kühberger, we consider the do-it-both-ways rule (which in the present context may be better called the do-it-n-ways rule) an investment that promises high payoffs. Waiting for a theory of hypothetical and real decisions to emerge from an empirical vacuum seems overly optimistic. In fact, a comprehensive collection of reliable effects of financial incentives (as would quickly evolve if the do-it-both-ways rule was applied in psychology and economics) may act as a strong incentive to develop such a theory.

5.3. Do incentives eliminate anomalies? From the 1970s, psychologists conducting research in the tradition of the heuristics-and-biases program have accumulated experimental evidence that suggests "behavioral assumptions employed by economists are simply wrong" (Grether 1978, p. 70). One prominent response of economists to this challenge has been to question the validity of the evidence. Experimentally observed anomalies ("fallacies," "biases") could be, so the argument goes, peculiar to the methodological customs and rituals of psychologists (e.g., Grether 1980, 1992; Grether & Plott 1979). A number of commentators (e.g., **Fantino & Stolarz-Fantino, Gil-While, Holt, Kühberger, Zizzo**) continue this debate and discuss the robustness of anomalies. **Hilton,** for instance, asks whether anomalies can be eliminated by financial incentives and learning, and he concludes that "the evidence strongly suggests that they cannot."

The picture is more differentiated and we agree with **Smit**h that more "constructivity" is needed. It is time to go beyond blanket claims and categorical questions such as whether or not financial incentives eliminate anomalies. We also agree with **Fantino & Stolarz-Fantino's** conclusion that the impact of methodological key variables on the results obtained "are not fixed and should not be taken for granted." **Holt & Laury's** study is a good example of how differentiated the

empirical pattern can be. In addition, it is important to acknowledge that conclusions regarding the effects of financial incentives (and similarly, repetition and feedback) are based on small and (sometimes opportunistic) samples of studies, and thus very likely are not the last word.

The little we know, however, suggests that financial incentives matter more in some areas than others (see Camerer & Hogarth 1999). Moreover, as we pointed out in Hertwig and Ortmann (in press), they matter more often than not in those areas that belong to the home turf of psychologists, namely studies on judgment and decision making. Ironically, they may matter less in "game and market experiments" (Camerer & Hogarth 1999, but see Smith & Walker 1993 and Schotter & Merlo 2000), the home turf of economists. The need for a theory of the effects of financial incentives is apparent. We suggest that Camerer and Hogarth (1999) is an excellent point of departure. Most importantly, these authors' capital-labor framework of cognitive effort highlights the interaction effects between key design variables such as repetition and financial incentives. Thus, it may have the potential to account for the heterogenous observations regarding the effects of financial incentives.

Related to the existence of anomalies, **Gil-White** asks "if (1) people are given rewards for being good Bayesians, and (2) they receive feedback that is immediate and highly accurate, should we -- upon the observation of results consistent with Bayesian reasoning -- conclude that we have shown that people are good Bayesians, or that experiments set up in this way can train them to be such?" The "context dependency of results" (**Levine**) is, indeed, an important question - but it is one that deserves equal attention in studies omitting rewards and feedback. To rephrase Gil-White: If people are given no rewards, and if they have only one or a few chances for an answer in an unfamiliar context, should we -- upon the observation of results inconsistent with Bayesian reasoning -- conclude that we have shown that people are bad Bayesians? In our view, Hogarth answers Gil-White's challenge**:** He reminds us that when faced with context-dependent results, researchers need to theoretically clarify the conditions to which results can be generalized.

Finally, we stress that the robustness of "biases", "cognitive illusions", or "fallacies" is not only debated in economics but also in psychology. In the wake of this debate, rich empirical and theoretical work in psychology has evolved that attempts to explain when and why people's inferences obey or disobey certain normative standards (see e.g., Erev, Wallsten & Budescu 1994; Hilton 1995; Gigerenzer 1996; Krueger 1998; Juslin, Winman, & Olsson 2000). Thus, focusing merely on learning and financial incentives overlooks, for instance, what is likely to be the most powerful tool to reduce and sometimes even to eliminate "blunders of probabilistic reasoning" (Tversky & Kahneman 1987), i.e., to present statistical information in terms of (natural) frequencies rather than probabilities (e.g., Gigerenzer 1991a; Gigerenzer & Hoffrage 1995; Cosmides & Tooby 1996; Hoffrage, Lindsay, Hertwig, & Gigerenzer 2000). The seemingly robust conjunction fallacy (that **Zizzo** discusses), for instance, can be reduced and sometimes completely eliminated when the information is presented in terms of frequencies (see Hertwig & Gigerenzer 1999, but also Mellers, Hertwig & Kahneman in press). Unfortunately, recent reviews of psychological literature for economists seem to be blissfully unaware of these empirical findings and theoretical discussions (e.g., see Rabin 1998).

6. Honesty versus deception

The use of deception in experiments entails costs. In light of the still frequent use of deception in some areas of psychology, it was surprising for us to see that none of the commentators explicitly question this assessment, and some commentators explicitly agree with it (e.g., **Baron**, **Goodie**, **McKenzie & Wixted, Zizzo**). We first clarify the definition of deception, then discuss McKenzie and Wixted's illustration of how suspicion contaminates experimental results, and finally explore when and why some commentators consider deception to be necessary.

6.1. Reprise: What is deception, and what is not? **Baron** proposes that not informing participants of the purpose of the study is "deception by omission"- but is it? In contrast to **Baron**, most researchers do not seem to regard the withholding of information as deception. Such agreement is, for instance, manifest in studies that review how often deception is used in psychological experiments. In Hertwig and Ortmann (2000), we examined the criteria for deception in those review studies. Intentional and explicit misrepresentation, that is, provision of false information, is unanimously considered to be deception. In contrast, not acquainting participants in advance with all aspects of the research being conducted, such as the hypotheses explored (e.g., the relationship between positive rewards and mood to use Baron's example) is typically not considered deception. This view is also shared across disciplinary boundaries as the following statement by John Hey (1998) illustrates: "There is a world of difference between not telling subjects things and telling them the wrong things. The latter is deception, the former is not" (p. 397).

Despite such a consensus, we appreciate **Baron**'s argument that withholding information makes full disclosure, when it is desirable, appear suspicious. In Ortmann and Hertwig (2000), we argued that one specific kind of "deception by omission" has the same potential for creating distrust as providing false information, namely, the violation of participants' default assumptions. For instance, a default assumption participants are likely to have is that a study starts only after an experimenter has clearly indicated its beginning. As a consequence, a participant might assume that her or his initial interactions with the experimenter (upon entering the laboratory) are not part of the experiment, and might feel mislead if she finds out otherwise. We propose that violating default assumptions should be avoided.

6.2. How participants' suspicions systematically contaminate data. **McKenzie & Wixted** provide two intriguing examples of how participants' distrust (likely to be fueled by the use of deception) systematically contaminated experimental results. Specifically, they show that the failure to recognize that participants may distrust experimenters' assertions about important task parameters (e.g., that a particular piece of information was randomly drawn) can lead participants' responses to be misclassified as irrational (e.g. as non-Bayesian). McKenzie & Wixted's analysis shows that participants' distrust has the potential to add random noise to the observations and can cause experimenters to draw erroneous conclusions - for instance, regarding people's ability to reason in accordance with normative principles (e.g., Bayes's rule). The threat of systematic contamination due to distrust has also been documented in other domains. In an extensive search for studies exploring the contaminating effects of deception, Hertwig and Ortmann (2000) and Ortmann and Hertwig (2000) found that, across a variety of research domains, personal experience with deception can and does distort observed behavior (e.g., judgments, attitudes, and measures of incidental learning and verbal conditioning).

As an option to deal with participants' skepticism about task parameters, **McKenzie & Wixted** propose incorporating a "trust" (or distrust) parameter into descriptive models of participants' behavior. While elegant, this approach introduces a free parameter into the models (thus increasing the danger of data-fitting, in particular when unlike in McKenzie and Wixted's models, more than one free parameter is involved). Moreover, we fear that this modeling approach will often not be applicable as it demands a good understanding of where and how distrust interferes with the processes under consideration.

6.3. Is deception indispensable, and is it treated as a last-resort strategy? The most common argument for deception suggests that it is indispensable for the study of those facets of behavior for which participants have reasons to conceal their truthful opinions, attitudes, or preferences. Therefore, experimenters must lie in order to avoid being lied to. Several commentators reiterated this argument (e.g., **Davis & Durham; Van Vugt**) or variants of it (**Baron**). In the absence of strong incentives to develop alternatives to deception, this rationale can only be evaluated in the abstract. Clearly, at this point there is no principled argument that could prove this rationale wrong. Consequently, we did state in the target article that we do not exclude the possibility that there are important research questions for which deception is indispensable. Irrespective of this issue, however, we argue that the prevalence of deception could substantially be reduced if it were used as a strategy of last resort. It is here where we disagree with Weiss - who whole-heartedly defends the current practices.

Weiss complains that our focus on the methodological costs of deception is an ethical argument in disguise. In suspecting an ethical argument, Weiss refers to our public good analysis of deception. Clearly, our argument is orthogonal to the deontological arguments put forth by Baumrind (1964, 1979, 1985). We agree with Weiss that our an analysis invites a value judgment, namely, that the experimenters who produce the public good while others do not are being exploited. But, such a judgement surely does not make the analysis of deception in terms of a public good problem less valuable; nor does it absolve the defenders of deception from discussing its contaminating potential. Weiss does not devote a single word to this potential. Ironically, Weiss himself points out that deception comes at a considerable cost, namely, that of an arms race in which experimenters have to design even more sophisticated ways of camouflaging the true purpose of the experiment since participants may become increasingly distrustful and sophisticated in figuring out where deception occurs[3].

Weiss compares the use of deception to the use of antibiotics to combat bacterial infection. We appreciate the analogy. To the great dismay of physicians and, in fact, all of us who are (future) patients, antibiotics have increasingly become ineffective. Why? Because, and for us this is the crucial lesson of Weiss's analogy, they are not used as last-resort treatment but instead as first-resort treatment in cases of common and uncomplicated infections. The unfortunate consequence is that bacterial resistance to antibiotics is more likely to evolve. For instance, bacterial resistance in the Netherlands is about 1%, compared with the US average of around 25%. In the US, it is routine to use antibiotics for treating middle ear infection, one of the most common diagnoses in children, whereas in other countries, such as the Netherlands, the standard practice is to use "watchful waiting" for one to two days after the onset of an ear infection in children over 2 years of age, treating only if the infection fails to improve during that time (see

http://healthlink.mcw.edu/article/965945751.html).
As medical researchers have to continuously invent new antibiotics to keep abreast of the microbes, behavioral researchers may need to create novel deception to stay ahead of participants' suspicions.

Just as antibiotics, deception is not exclusively used as a last-resort strategy. In contradiction to **Davis & Durham's** belief, even a cursory glance at contemporary deception studies reveals that deception is used even when it is not indispensable (recall that every third study in JPSP and every second study in JESP uses deception). Instances are the claim that incentives are performance contingent when in reality they are not or the claim that some aspect of an experiment (e.g., the allocation of a specific role, a piece of information, etc.) was randomly chosen when in reality it was not (for more details see Hertwig & Ortmann in press). Deception is a method that saves resources (as **Baron** points out) but it is only inexpensive if there will be no costs for future experiments. But are there really no costs? We doubt that the participants believe promises of performance-dependent rewards at face value in future experiments if they just found out (through debriefing) that the experimenter misled them on the contingency of those rewards. Once bitten, twice shy.

The evidence regarding the consequence of firsthand experience with deception (see Hertwig & Ortmann 2000; Ortmann & Hertwig 2000) counsels us to treat deception as a last-resort strategy, thus limiting the number of participants with firsthand experience. In fact, this is the policy as currently stipulated by the APA guidelines. Considerations as formulated by **Baron** (resource savings) and **Van Vugt** (deception is justified when beneficial in the development of non-trivial theories of human behavior) are not endorsed by the APA guidelines. Finally, experiments that a number of commentators (e.g., **Davis & Durham, Hilton, Roth**) consider to be prime examples of cases in which deception was indispensable or yielded valuable, and eminently teachable, insights into human behavior would likely not pass contemporary ethical review committees (e.g., the Milgram experiment). Therefore, those studies do not help much in inferring the costs and benefits of future deception studies.

How can deception be implemented as a last-resort strategy and how can the existing distrust among participants be overcome? In Hertwig and Ortmann (2000), we propose an incentive-compatible mechanism that has the potential to reduce deception (and to promote methodological innovation). To deal with the effects of existing distrust, each individual laboratory can attempt to (re-)establish trust by taking measures such as introducing a monitor into experimentation (i.e., participants elect one of themselves to be a paid monitor who inspects all equipment and observes all procedures during the experiment; see Grether 1980, 1992; for a similar proposal see **Baron**). Such concrete gestures to (re-)gain participants' trust may also help to shorten the time any policy change will require for psychologists to overcome their reputation (a problem that **Roth** points out).

As still another remedy, **Henrich** proposes conducting deception studies outside the laboratory. Albeit an original proposal, we doubt its long term utility. If psychologists restricted the use of deception to field studies, this practice would quickly become public knowledge. Such knowledge, and the expectations it is likely to evoke, may compromise not only the work of researchers who conduct field studies but also that of professional psychologists in general.

7. Additional issues: Participant pools, institutional arrangements, and the art of fast data collection

In reflecting on how we go about our business, several commentators highlighted practices that, in their and our view, deserve closer scrutiny. **Henrich**, for instance, criticizes the reliance of both psychology and economics on university students as participants - "a very weird, and very small, slice of humanity." He argues that as a result of this practice, researchers from both fields overgeneralize their results.

Psychology's reliance on a highly selected subject pool may be even more pronounced than Henrich assumes. According to Sieber and Saks, "undergraduate students have been a major source of research data in many areas of psychology" (p. 1053). The participation of undergraduates in the subject pool is typically institutionalized through the requirement that students in introductory psychology need to participate in (some) research projects as part of their course requirements. This availability of "free subjects" may be a key to understanding psychologists' experimental practices. Vernon Smith (personal communication) once asked a psychologist "why psychologists, who you would expect to be concerned about motivation, did not pay salient rewards to subjects. [The psychologist] said it was simple. Every psychology department requires majors to participate in a minimum number of experiments as a condition for a degree, and that it was unthinkable that you would start using rewards for this huge population of free subjects."

There is evidence indicating that since the 1960s the proportion of participants in psychology experiments from introductory courses has been on the rise (see Hertwig & Ortmann in press). One way to read this change is that the current widespread recruitment from introductory courses is an institutional response to the risks of a subject pool contaminated by distrust (Ortmann & Hertwig, 2000). Recruiting participants who are less likely to have firsthand experience with deception - students in introductory classes - minimizes the problem of participants' suspicions. Albeit speculative, this explanation is consistent with the advice psychologists were given not to use the same (deceived) students twice. Based on experimental evidence, Silverman, Shulman and Wiesenthal (1970) concluded, more than 30 years ago, "that the practice of using the same subjects repeatedly be curtailed, and whenever administratively possible, subjects who have been deceived and debriefed be excluded from further participation" (p. 211). Interestingly, and supporting our argument about the detrimental consequences of deceptive practices in psychology, increased reliance on introductory students has- to the best of our knowledge – not been observed in economics**.**

In our view, **Hogarth** suggests one important step toward a systematic remedy of a potentially unrepresentative subject pool. He calls for theoretical clarity about the kinds of people and tasks to which the results might be generalized. In addition, he advocates the combination of laboratory and field studies, thus rendering it possible for experimenters to explore the generalizability of their laboratory results. In Hogarth's view, economists more than psychologists explicate the characteristics of people and tasks to which experimental results are meant to be generalized. By attending to the task, Hogarth reminds psychologists of a concern that is reminiscent of Egon

Brunswik's (1955). He criticized his colleagues for practicing "double standards" by generalizing their results to both a population of situations and a population of people, although only being concerned with the sampling from the latter.

Gigerenzer describes a tendency in psychological experimentation to conduct large-scale data collection outside the laboratory context, in particular, the practices of "take-home questionnaires" and "large classroom experiments." These methods are another instance of fast and seemingly inexpensive methods. As Gigerenzer suggests, they may, however, explain why different researchers observe divergent results in studies researching probabilistic and logical reasoning. Ultimately, those methods may also have contributed to why much past research in behavioral decision making has arrived at rather pessimistic conclusions regarding people's reasoning abilities. Again, the do-it-both-ways rule is a tool for studying to what extent experimental results are contingent on the methods of data collection (e.g., classroom studies vs. small-group laboratory experiments).

8. Why do the methodological practices differ?

A number of commentators argue that the answer simply is that practices (must) differ because the two disciplines have different subject matters and research questions (e.g., **Davis & Durham, Gil-White**, **Lecoutre & Lecoutre, Suleiman**). But is it "really the question being asked that will always determine the methodology" (**Davis & Durham**; see also for a similar form-ought-to-follow-function argument, **Baron, Huettel & Lockhead, Lecoutre & Lecoutre, Roth**)? **Gigerenzer** reveals the ahistorical naivite of this claim. Even within psychology, this "naturalistic" argument fails to explain surprising structural similarities in different theoretical traditions. What are more substantial explanations of the observed differences? **Blaich & Barreto** suggest that the different practices may be driven by economists' and psychologists' different use of statistics. **Gintis** reiterates the conjecture in our target article of the unifying role of game theory, arguing that the emergence of game theory suffices to explain the differences in psychologists' and economists' experimental practices. Finally, **Huettel & Lockhead** argue that psychologists' and economists' experiments serve different functions.

Before proceeding to address these proposals, let us distinguish two aspects of methodological standards, namely their 'nature' and their 'content.' The former refers to how binding standards are; the latter to their actual substance (e.g., use of financial incentives). We suggest that any explanation of the different nature of standards in economics and psychology needs to acknowledge the broader context from which experimentation emerged in either field. **Hogarth** and **Roth** reiterate our suggestion that experimental economists had to fight hard to be accepted in a profession that for many years doubted the utility of laboratory experiments for making inferences on the real world (as can be glimpsed from the introductory remarks of two path-breaking papers, Smith 1976, 1982). In contrast, experimental psychologists never similarly had to battle for respect within their own discipline. While the specific circumstances of their emergence may explain why methodological standards are less binding in psychology than in economics, history does not explain the content of the standards. How did the content evolve?

8.1. Use of statistics. In **Blaich & Barreto's** view, research practices differ because psychologists and economists make different use of inferential statistics: "The fact that experimental psychologists tend to assign much more importance to rejecting the null hypothesis but less importance on making precise parameter estimates than experimental economists plays an important rule, in our view, in creating the differences in the two fields." Although we are not convinced that these differences have caused different methodological practices (e.g., psychologists use of deception and economists proscription of deception), it may very well be that psychology's practice of null-hypothesis testing perpetuates differences. It does so by impeding the elaboration of precise theories (an argument that has repeatedly been made within psychology, Schmidt & Hunter 1997; Krueger 1998; Hertwig & Todd 2000). Imprecise theories, in turn, tend to leave decisions on how to realize experimental tests to the discretion of the researchers, and thus to the dominant preferences in a field. In contrast, precise theories are more likely to imply appropriate test conditions, for instance, by explicitly defining the behavior it targets (e.g., first impression, learning, equilibrium behavior).

8.2 The role of game theory. Gintis claims that there is a simple answer as to why methodological practices differ. It is because economists use game theory to design and interpret experiments. Although we hinted at the unifying role of game theory, its presence cannot explain why methodological conventions have such a regulatory nature in economics. We believe that the most plausible candidate explanation for their nature is the strategic role that the canonization of mandatory rules played in the process of persuading the discipline. With regards to the content of the conventions, **Gillies & Rigdon** argue - contrary to **Gintis's** thesis - that three of the four key variables (namely, deception, financial incentives and scripting) are "general" variables "whose appropriateness is independent of the theory being tested." Whether or not this assessment is correct, we anticipate that any comprehensive explanation of why methodological practices in the two fields differ will encompass disparate roots - among them the role of early key players (**Holt & Laury** and **Smith** point out that a psychologist, Sidney Siegel, has been largely responsible for establishing the procedural standards used in economics experiments), the role (or relative lack thereof) of unifying theories (e.g., game theory, behaviorism), and institutional arrangements (e.g., the availability of subject pools) as well as the fact that experimental economics for a significant number of years was done only in about half a dozen laboratories.

While discussing game theory, let us note the assertion that "economists [and hence experimental economists] use the theory of self-interest as unique explanatory framework for understanding human behavior," (**Van Vugt**) is wrong. It demonstrates plain unawareness of the theoretical developments that have dramatically reframed - by way of mostly game-theoretic reformulations - economics. We doubt that there are economists out there who do not believe that "given the right conditions, people can be rational or irrational, selfish or altruistic, aggressive or helpful" (Van Vugt). We are certain that if indeed such an "econ" (Leijonhufvud) exists, he or she is not an experimental economist. This author's wage escalation argument, furthermore, misapplies basic tenets of marginal utility theory. Money is typically chosen because of its, for all practical purposes, non-satiation property. By Van Vugt's logic real wages would go up and up and up ... . Last but not least, and also regarding issues of homo oeconomicus, **Zizzo** takes us to task for our comments on Frey's work on intrinsic motivation. We urge the reader to re-read footnote 8 of the target article and read the references therein.

8.3. Do psychologists generate questions, whereas economists test models? **Huettel & Lockhead** make a distinction between "restricted experimental designs which allow reproducibility and hypothesis testing, and exploratory designs, which provide new insight into phenomena." Based on this distinction, they suggest that economics studies were designed to answer well-defined hypotheses, whereas the psychology studies in question have more of a exploratory character.

Huettel & Lockhead's characterization of the psychological studies in question is not well-informed. Research in experimental economics and psychological research on judgment and decision making are particularly well suited for a comparison of methods across disciplines because studies in both fields often address similar and sometimes even identical questions. As examples, consider questions such as whether or not people update probabilities in a Bayesian way, make choices in a transitive way, are subject to the hindsight bias ('curse of knowledge'), or allocate resources in a way that satisfies rational economic theory (or motives such as fairness). Sometimes economists and psychologists explore exactly the same hypothesis: for instance, whether or not people apply the representativeness heuristic (Kahneman & Tversky 1973) to update probabilities (e.g., Grether, 1980, 1992; Harrison 1994). Arguing, as Huettel & Lockhead do, that the economists' and "the particular psychology studies that were selected for consideration" differ because the latter are in the business of "generating questions," whereas the former test well-defined hypotheses reveals, depending on the perspective, a rather self-depreciating or condescending attitude. We do, however, agree with Hogarth's assessment that many psychological studies test theoretical notions rather than formalized theories or process models - exactly this fact has been at the heart of a controversial debate among psychologists (Kahneman & Tversky 1996; Gigerenzer 1996).

Despite our disagreement with how **Huettel & Lockhead** characterize the goal of psychological studies, we appreciate the more general question they raise, namely, whether or not our policy recommendations should (equally) apply to hypothesis-testing and hypothesis-generating experiments. While we agree with Huettel & Lockhead that in the context of discovery "everything goes", we point out that placing participants in a not well-defined situation is only one and probably not a particularly productive tool to generate questions. In the context of theory testing, **Erev** highlights one of the crucial benefits of standardized test conditions, namely the emergence of data sets that, because of being collected under comparable conditions, can be used in toto to test a hypothesis, a model, or a theory. Such a body of data will allow researchers to use a strict rule, "the generality first" rule, in the process of theory selection. This rule requires that a new model replaces an old model only if it explains previous data plus new data that the old model cannot accommodate. While we suggest that this rule should not be used in isolation (but be complemented by other theory-selection criteria such as internal consistency and simplicity), we agree with Erev that the evolution of large standardized data sets is one promising route to cumulative progress in modeling. We also agree with Erev that the do-it-both-ways rule will first quickly help to identify how key variables of experimental design affect the results obtained, and then, once such knowledge is available, will promote the evolution of data sets collected under comparable conditions. As an aside, the generality-first rule also implies a third way of testing theories -- beyond null-hypothesis testing and parameter estimation (**Blaich & Barreto**). According to this rule, a theory is tested against the aggregate set of data, and its status (rejected/accepted) is a function of its explanatory power (regarding this set) and the performance

of its competitors. Because they are intended to be useful approximations (Roth), theories can overcome rejection based on individual experiments if they still succeed in accounting for a wide range of observations.

9. Experimental economics and behaviorism: Guilty by association?

> "I have not the slightest doubt that if Sid Siegel had lived, say another 25-30 years, the development of experimental economics would have been much advanced in time. He was just getting started, was a fountain of ideas, a powerhouse of energy, and had unsurpassed technique and mastery of experimental science. 25 years later I asked Amos Tversky, 'Whatever happened to the tradition of Sidney Siegel in psychology.' His answer: 'YOU'RE IT!'"

Communicating this episode to us, Vernon Smith made it clear that this comment was meant to be a put-down because Tversky saw Siegel as one of the last of the Skinnerian behaviorists. Likewise, a number of commentators noted the similarities between the experimental practices of economists and those of experimental psychologists in the learning tradition. **Fantino & Stolarz-Fantino** see this similarity in a positive light and illustrate how classic effects observed in the heuristics-and-biases program (e.g., base-rate neglect, conjunction fallacy) can be studied using methods from the learning tradition. For **Hilton** and **Kurzban** (and in a somewhat related way also **Maratsos**) in contrast, this similarity is a reason for concern. Admittedly simplified, Hilton and Kurzban's arguments are the following: First, the experimental methods in economics resemble those employed by behaviorists. Second, the methodological similarity indicates a theoretical affinity, with economists being "methodological behaviorists" who focus on observables at the expense of cognitive processes (Hilton; see also **Markman, Rakow**), or focus, like behaviorists do, on domain-general non-significant learning mechanisms. Third, either focus is a theoretical cul de sac, and "psychologists did the right thing to abandon behaviorism" (Hilton), whereas adopting economists' methodology in psychology would be tantamount to "behaviorist-like experiments" and akin to a return to the "dark days of behaviorism" (Kurzban).

We disagree with **Hilton's** and **Kurzban's** view. They seem to suggest that taking into account realizations of key variables of experimental design that economists and behaviorists value, goes along with adopting their imputed theoretical biases (i.e., focus on output or nonsignificant learning mechanisms). As Gigerenzer explains, however, there is no such automaticity. Even within psychology, there are research programs that are utterly different in their theoretical nature despite commonalities in experimental practices. Thus, even if it were true that economists focus on observable outcomes (at the expense of processes) as Hilton and **Rakow** suggest, nothing in the emphasis on learning and motivation excludes the study of processes [as, for instance, Wallsten's 1972, 1976, studies on probabilistic information processing illustrate]. On the contrary, the provision of financial incentives is one important tool for decreasing variability, thus increasing the reliability of processes; and the use of repeated trials is the tool to study the evolution of processes. There is hardly an automatic contingency between the use of financial incentives, scripts, and repetition, and the sudden disappearance of cognitive processes in a black

box.

But is the conventional wisdom that **Hilton** and **Lecoutre & Lecoutre** express even accurate - that psychologists are process-oriented, whereas economists focus on observable outcomes? There are important counterexamples. Take, for instance, the most influential research program in psychological research on behavioral decision making, the heuristics-and-biases program. It seems not unfair to conclude that this program has an explicit focus on observable outcomes (**Markman** seems to agree). Compared to the search for new "biases", "fallacies", and "cognitive illusions", the modeling of the psychological processes has received little attention [see the debate between Kahneman & Tversky (1996) and Gigerenzer (1996)]. Take also research programs in economics signified by Camerer et al. (1993) or Costa-Gomes, Crawford & Broseta (forthcoming), or McCabe et al. (2000; see also **Smith**). Whereas these researchers are still interested in outcomes, they focus on the reasoning processes underlying choices leading to outcomes, and even their neurological correlates.

Finally, what about **Kurzban's** argument that economists (and behaviorists' alike) study domain-general mechanisms of non-significant relevance? Although we are not sure what mechanisms Kurzban has in mind, it is worth remembering that theorizing about domain-specificity (as evolutionary psychologists such as Kurzban do) apparently can profit from domain-general theoretical frameworks such as game theory. Take Cosmides and Tooby's (1992) social contract theory, one major theory in recent evolutionary psychology, as an example. In their view, the barrier to the evolution of social exchange is a problem that is structurally identical to the one-move Prisoner's Dilemma, and indeed Cosmides and Tooby (1992) used this game to refine their theory.

10. Experimental practices in psychology: A challenge for economists?

We whole-heartedly agree with the opinion of those commentators who point out that the methodological dialogue between economists and psychologists must not be a one-way street (**Harrison & Rutström, Levine, Roth**). Indeed, such a debate can only work if both sides are open to the input from the other camp. Admittedly, we focused in our treatment on those key variables where we believe psychologists can profit from comparing their experimental practices with those of experimental economists, the new kid on the block.

We also pointed out, however, that the conventions and practices of experimental economists do not constitute the gold standard of experimentation, and that "a paper entitled 'Experimental practices in psychology: A challenge for economists?' may well be worth writing." To our minds, there is no doubt that **Harrison & Rutström** are right on the money when they argue that "unfortunately, experimental economists have sometimes followed conventional practices with little thought about the consequences." We share their scepticism of "the popular use of 'lab dollars'" (as, incidentally, do Davis & Holt 1993, p. 29). More generally, we also stress that the do-it-both-ways rule is a significant departure from empirically blind conventions that experimental economists currently take for granted.

10.1. Acontextualization, field referents, and framing effects**.** **Harrison & Rutström** also discuss the issue of field referents that participants may bring into the laboratory. Economists typically try to overcome the problem of such imported priors by acontextualization – stripping the experimental scenario and instructions of any reference to the real-world problem that may have motivated the scenario. For example, in principal-agent games most experimental economists label the employee the "seller" and the employer the "buyer" of unspecified goods or services. Sometimes they even omit these labels and call the employee (employer), say, "participant A" ("participant B"). Although acontextualization has the advantage of counteracting the problems of uncontrolled priors that participants bring into the laboratory (an issue that **Fantino & Stolarz-Fantino**, **Goodie**, and **Betsch & Haberstroh** also highlight), it has two clear drawbacks. First, the abstract context invites sense-making exercises on the part of the participants who might try to make the connection between the laboratory set-up and possible real-world correlates. Second, the abstract context may prevent participants from invoking the kind of inference routines that they use to navigate similarly structured real-world environments. We use the word "routines" here intentionally because, although we disagree with their claim about the scope, we accept Betsch & Haberstroh's basic point about the importance of automatic processes.

Relatedly, **Hogarth** argues that "theory (in economics) specifies that different structural representations of the environment (e.g., framing of decision problems) should make no difference. .... Context - however vaguely defined - is important to psychologists, but not to economists." Although that statement is not true in its generality (e.g., Andreoni 1995; Offerman, Sonnemans, Schram forthcoming; Ortmann, Boeing, Fitzgerald 2000; or the previously mentioned work by **Smith** and his collaborators), there can be no doubt that psychologists are overwhelming more sensitive to how problem and information representation affects people's reasoning.

10.2. Heuristics and how to select them. **Harrison & Rutström** highlight the selection of heuristcs as a theoretical theme that unites the field of experimental economics and psychology.

We agree whole-heartedly. More generally, we believe that in a world in which knowledge and mental resources are limited, and in which time is pressing, the study of real-world judgments and decisions require alternatives to traditional models of unbounded rationality and optimization. In a recent BBS précis, Todd and Gigerenzer (2000) described the framework of fast and frugal heuristics and placed the study of those heuristics within the context of bounded rationality (Simon 1991). Moving beyond the description of heuristics, Todd, Gigerenzer & the ABC Research Group (2000) delineate three major theoretical questions: Where do heuristics come from? How are they selected and how are they adapted to the decision and environment structure in which they evolved? Seeking to answer these and related questions can foster a further theoretical convergence. Although we are not sure that the search for boundedly rational heuristics is what **Levine** envisions when he talks about a "psychologically based economic theory," we agree with him that – whether they function as stopping rules for search (Simon 1956) or in some other way -- will be an crucial topic in any program of bounded rationality.

Theoretical and methodological issues are often linked. The study of heuristics is a case in point. Obtaining empirical evidence for the use of particular heuristics demands careful methodology because of challenges such as the flat maximum phenomenon (see Todd & Gigerenzer 2000) and individual differences in their use (a source of variance that **Suleiman** stresses). The study of heuristics will require psychologists and economists to make methodological decision closely related to those that we have discussed here -- decisions about the structure of the decision environment (e.g., abstract vs. content-rich), the incentive landscape (e.g., favoring accuracy, speed, or other performance criteria), or the structure and kind of feedback (to study the evolution and learning of heuristics). We agree with **Markman** that psychology has much to offer in terms of techniques for studying on-line processing and heuristics. In fact, as pointed out above, economists have already started using techniques such as MouseLab. Smith reminds us correctly that methodological and ultimately theoretical progress is intimately linked to the technological tools with which we produce knowledge (see also Gigerenzer 1991b)

11. Conclusion

> "Methodological discussion, like spinach and calisthenics, is good for us ... ."
> (Paul Samuelson)

Commenting on presentations by Ernest Nagel, Sherman Krupp, and Andreas Papandreou on methodological problems, Samuelson (1963) noted that while undoubtedly methodological discussion is good for us, it is – like spinach and calisthenics – not often practiced and thus may be, ultimately, inconsequential. **Roth** argues similarly that "because experiments are part of scientific conversations that mostly go on within disciplines, differences in standard practices between disciplines are likely to persist." We hope that he is wrong and that **Harrison & Rutström** are right with their generous claim about the effect of our target article. There is hope. After all, more people today appreciate spinach and calisthenics (although they typically have fancier names for the latter).

We are convinced that a common language of scientific discovery and theory-testing, in addition to experimental practices grounded in empirical evidence, promise high payoffs. Ultimately, of

course, these claims are an empirical question. We can say for ourselves – one being a psychologist, the other being an economist – that we found the conversation across disciplinary borders a rewarding (albeit not always easy) exercise. We urge others to follow suit.

References

* = Starred references were already used in the target article.

Andreoni, J. (1995) Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. Quarterly Journal of Economics 110:1-21.

Bardsley, N. (in press) Control Without Deception: Individual Behaviour in Free-Riding Experiments Revisited. Experimental Economics. Experimental Economics.

Barron, G. & Erev, I. (2000) On the relationship between decisions in one-shot and repeated tasks: Experimental results and the possibility of general models. Manuscript, Faculty of Industrial Engineering and Management, Technion.

*Baumrind, D. (1964) Some thoughts on ethics of research. After reading Milgram's Behavioral study of obedience. American Psychologist 19:421-423.

*Baumrind, D. (1971) Principles of ethical conduct in the treatment of subjects: Reaction to the draft report of the Committee on Ethical Standards in Psychological Research. American Psychologist 26:887-896.

*Baumrind, D. (1979) IRBs and social science research: The costs of deception. IRB: A Review of Human Subjects Research 1:1-4.

*Baumrind, D. (1985) Research using intentional deception: Ethical issues revisited. American Psychologist 40:165-174.

Berninghaus, S.K. & Ehrhart, K.M. (1998) Time horizon and equilibrium selection in tacit coordination games: Experimental results. Journal of Economic Behavior & Organization 37:231-248.

*Binmore, K. (1994) Playing fair. The MIT Press.

*Binmore, K. (1999) Why experiment in economics? The Economic Journal 109:16-24.

Blume, A. & Ortmann, A. (2000) The effects of costless pre-play communication: Experimental evidence from a game with Pareto-ranked equilibria. Manuscript submitted for publication.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. Psychological Review 62:193-217

Burnham, T., McCabe, K. & Smith, V.L. (2000) Friend-or-foe intentionality priming in an extensive form trust game. Journal of Economic Behavior & Organization 43:57-73.

*Camerer, C. F. (1997) Rules for experimenting in psychology and economics, and why they differ. In: Understanding strategic interaction. Essays in honor of Reinhard Selten (pp. 313-327),

ed. W. Albers, W. G¸th, P. Hammerstein, B. Moldovanu & E. von Damme. Springer.
*Camerer, C., Loewenstein, G. & Weber, M. (1989) The curse of knowledge in economic settings: An experimental analysis. Journal of Political Economy 97:1232-1255.

Camerer, C.F. & Hogarth, R.M. (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. Journal of Risk and Uncertainty 19:7-42.

Camerer, C., Johnson, E., Rymon, T., & Sen, S.(1993) Cognition and framing in sequential bargaining for gains and losses. In: Frontiers of Game Theory (pp. 27-47), (eds. K. Binmore, A. Kirman, & P. Tani . MIT Press.

Charness, G. & Rabin, M. (2000) Social preferences: Some simple tests and a new model. Working paper, Working paper, Department of Economics, University of California-Berkeley.

Cosmides, L, & Tooby, J. (1992) Cognitive adaptations for social exchange. In: The adapted mind: Evolutionary psychology and the generation of culture (pp. 163-228), eds. J. H. Barkow, L. Cosmides & J. Tooby. Oxford University Press.

*Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition 58:1-73.

Costa-Gomes, M., Crawford, V. & Broseta, B. (forthcoming) Cognition and Behavior in normal-form games: An experimental study. Econometrica.

*Davies, M. F. (1992) Field dependence and hindsight bias: Cognitive restructuring and the generation of reasons. Journal of Research in Personality 26:58-74.

*Davis, D. D. & Holt, C. A. (1993) Experimental economics. Princeton University Press.

Dufwenberg, M. & Kirchsteiger, G. (2000) A Theory of Sequential Reciprocity. Working paper, Center for Economic Research, Tilburg University.

Edwards, W. (1954) The reliability of probability preferences. American Journal of Psychology 67: 68-95.

Erev, I., Wallsten, T. S., & Budescu, D.V. (1994) Simultaneous over- and underconfidence: The role of error in judgment processes. Psychological Review 101:519-527.

Falk, A. & Fischbacher, U. (1999) A theory of reciprocity. Working paper, Institute for Empirical Research in Economics, University of Zurich.

Gigerenzer, G. (1991a) How to make cognitive illusions disappear: Beyond heuristics and biases. In: European review of social psychology (vol 2, pp. 83-115), eds. W. Stroebe & M. Hewstone. Wiley.

Gigerenzer, G. (1991b) From tools to theories: A heuristic of discovery in cognitive psychology.

Psychological Review 98, 254-267.
*Gigerenzer, G. (1996) On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). Psychological Review 103:592-596

*Gigerenzer, G. & Hoffrage, U. (1995) How to improve Bayesian reasoning without instruction: Frequency Formats. Psychological Review 102:684-704.

*Goeree, J. K. & Holt, C. A. (in press) Stochastic game theory: For playing games, not just for doing theory. Proceedings of the National Academy of Sciences.

Goeree, J. & C.A. Holt (2000) A model of noisy introspection. Working paper, Economics Department, University of Virginia.

Goeree, J. & C.A. Holt (forthcoming) Ten Little Treasures of Game Theory, and Ten Intuitive Contradictions. American Economic Review.

Goodie, A.S., Ortmann, A., Davis J.N., Bullock, S., & Werner, G.M. (1999) Demons Versus Heuristics in Artificial Intelligence, Behavioral Ecology, and Economics. In: Simple heuristics that make us smart (pp. 327-355), eds. G. Gigerenzer, P.M. Todd, & the ABC Research Group. Oxford University Press.

Grether, D. M. (1978). Recent psychological studies of behavior under uncertainty. American Economic Review 68:70-74.

*Grether, D. M. (1980) Bayes rule as a description model: The representativeness heuristic. Quarterly Journal of Economics 95:537-557.

*Grether, D. M. (1992) Testing Bayes rule and the representativeness heuristic: Some experimental evidence. Journal of Economic Behavior and Organization 17:31-57.

*Grether, D. M. & Plott, C. R. (1979) Economic theory of choice and the preference reversal phenomenon. American Economic Review 69:623-638.

* Harrison, G. W. (1994) Expected utility theory and the experimentalists. Empirical Economics 19:223-253.

*Harrison, G. W. (1999) Experimental economics and contingent valuation. Working Paper 96-10, Division of Research, The Darla Moore School of Business, University of South Carolina (http://theweb.badm.sc.edu/glenn/eecv.pdf).

Henrich, J. (2000) Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. American Economic Review 90:973-979.

*Hertwig, R. & Gigerenzer, G. (1999) The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. Journal of Behavioral Decision Making 12: 275-305.

Hertwig, R.& Ortmann, A. (2000) Does deception destroy experimental control? A Review of the evidence. Manuscript submitted for publication.

Hertwig, R. & Ortmann, A. (in press). Economists' and Psychologists' Experimental Practices: How They Differ, Why They Differ, And How they Could Converge. In: Economics and psychology, eds. I. Brocas & J. D. Carillo. Oxford University Press.

Hertwig, R. & Todd, P.M. (2000) Biases to the left, fallacies to the right: Stuck in the middle withnull hypothesis significance testing. Commentary on Krueger on social-bias. PSYCOLOQUY 11(28). http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?11.028.

Hertwig, R & Todd, P. M. (in press). More is Not Always Better: The Benefits of Cognitive Limits. In: Reasoning and decision making: A handbook, eds. D Hardman & L. Macchi . Wiley.

*Hey, J. D. (1991) Experiments in economics. Basil Blackwell.

Hoffman, E., McCabe, K., & Smith, V.L. (2000) The impact of exchange context on the activation of equity in ultimatum games. Experimental Economics 3:5-9.

Hoffrage, U., Lindsey, S., Hertwig, R. & Gigerenzer, G. (2000) Communicating statistical information. Science 290:2261-2262.

Juslin, P., Winman, A., & Olsson, H. (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. Psychological Review 107: 384-396

*Kahneman, D. & Tversky, A. (1973) On the psychology of prediction. Psychological Review 80:237-251.

*Kahneman, D. & Tversky, A. (1996) On the reality of cognitive illusions: A reply to Gigerenzerís critique. Psychological Review 103:592-596.

Keren, G. & Wegenaar, W.A. (1987) Violation of utility theory in unique and repeated gambles. Journal of Experimental Psychology: Learning, Memory and Cognition 13:387-391.

Krueger, J. (1998). The bet on bias: A foregone conclusion? PSYCOLOQUY 9(46). ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/psyc.98.9.46.social-bias.1.krueger http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.46

McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T. (2000) A functional imaging study of "Theory of Mind" in two-person reciprocal exchange. Working paper, Economic Science Laboratory, University of Arizona.

*Mellers, B. A., Hertwig, R. & Kahneman, D. (in press). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. Psychological Science.

Offerman, T., Sonnemans, J., & Schram A. (forthcoming) Expectation formation in step-level public good games. Economic Inquiry.

Olcina, G. & Urbano, A. (1994) Introspection and equilibrium selection in 2x2 matrix games. International Journal of Game Theory 23: 183-206.

Ortmann, A. & Hertwig, R. (2000). The costs of deception? Evidence from psychology. Manuscript submitted for publication.

Ortmann, A., Fitzgerald, J., & Boeing, C. (2000) Trust, reciprocity, and social history: a re-examination. Experimental Economics 3: 81-100.

*Rabin, M. (1998) Psychology and economics. Journal of Economic Literature 36:11-46.

Rutstroem, E.E. (1998) Home-grown values and the design of incentive compatible auctions. International Journal of Game Theory 27: 427-441.

Samuelson, P.A. (1963) Discussion contribution. American Economic Review May 1963, vol 53.2. pp. 231 - 236.

Samuelson, P. & Nordhaus, W. (1985) Principles of economics (12th edition). McGraw-Hill.

Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: What If There Were No Significance Tests? (pp. 37-64), eds. L. L. Harlow, S. A. Mulaik, J. H. Steiger. Lawrence Erlbaum.

Schotter, A. & Merlo, A. (1999) A surprise-quiz view of learning in economic experiments. Games and Economic Behavior 28: 25-54.

*Sieber, J. E. & Saks, M. J. (1989) A census of subject pool characteristics and policies. American Psychologist 44:1053-1061.

Silverman, I., Shulman, A. D., & Wiesenthal, D. L. (1970) Effects of deceiving and debriefing psychological subjects on performance in later experiments. Journal of Personality and Social Psychology 14: 203-212.

Simon, H.A. (1956) Rational choice and the structure of environments. Psychological Review 63: 129-138.

Simon, H. A. (1990) Invariants of human behavior. Annual Review of Psychology 41:1-19.

Smith, A. (1976[1759]) The theory of moral sentiments. Oxford University Press.

Smith, Maynard J. (1982) Evolution and the theory of games. Cambridge University Press.

Smith, V. L. (1976) Experimental economics: Induced value theory. American Economic Review

Proceedings 66:274-279.

Smith, V. L. (1982) Microeconomic systems as an experimental science. American Economic Review 72:923-955.

Smith, V. L. & Walker, J. M. (1993) Monetary rewards and decision costs in experimental economics. Economic Inquiry 31:245-261.

Stahl, D. & Wilson, P. (1995) On players' models of other players: Theory and experimental evidence. Games and Economic Behavior 10: 218-254.

Todd , P.M.& Gigerenzer, G. (2000) Precis of Simple heuristics that make us smart. Behavioral and Brain Sciences 23:727-741.

Todd, P.M., Gigerenzer, G. & the ABC Research Group (2000) How can we open up the adaptive toolbox? Behavioral and Brain Sciences 23:767-780.

Tversky, A. & Kahneman, D. (1987) Rational choice and the framing of decisions. In: Rational choice: The contrast between economics and psychology (pp. 67-94), R. M. Hogarth & M. W. Reder (eds.). University of Chicago Press.

Van Huyck, J., Battalio, R.C., Mathur, S., Van Huyck, P., & Ortmann, A. (1995) On the origin of conventions: Evidence from symmetric bargaining games. International Journal of Game Theory 24: 187-212.

* Wallsten, T. S. (1972). Conjoint-measurement framework for the study of probabilistic information processing. Psychological Review 79:245-260.

*Wallsten, T. S. (1976) Using conjoint-measurement models to investigate a theory about probabilistic information processing. Journal of Mathematical Psychology 14:144-185.

Weibull, J.W. (1995) Evolutionary Game Theory. MIT Press.

Wundt, W. (1894). Lectures on Human and Animal Psychology. Translated from the second German edition by J.E. Creighton & E.B. Titchener. London: Swan Sonnenschein; the German edition is Wundt, W. (1892). Vorlesungen ueber die Menschen- und Thierseele. Hamburg und Leipzig: Leopold Voss.

Young, P. (1993) The Evolution of Conventions. Econometrica 61: 57-84.

1. We use the term do-it-both-ways rule as a short-hand expression. Obviously, there are situations where more than two realizations of a variable will be explored.

2. We note that there is one important exception to that statement: The work Cummings and Harrison and their collaborators have done on hypothetical bias in contingent valuation studies (see Harrison 1999 for an excellent survey and discussion).

3. Parenthetically, we note that we believe this concern to be significantly more relevant than Van Vugt's concern about participants being less likely to turn up again after having experienced deception once.